IDENTIFICATION OF MOLECULAR SIGNATURE IN EPITHELIAL TUMOURS

THESIS SUBMITTED TO
NATIONAL INSTITUTE OF TECHNOLOGY, ROURKELA
FOR THE PARTIAL FULFILMENT OF
THE MASTER DEGREE IN LIFE SCIENCE

Submitted By:

Subhrata Jena

Roll No- 411ls2048

Supervised By:

Dr. Bibekanand Mallick

Asst. Professor

DEPARTMENT OF LIFE SCIENCE
NATIONAL INSTITUTE OF TECHNOLOGY,
ROURKELA -769008

2013

NATIONAL INSTITUTE OF TECHNOLOGY, ROURKELA
राष्ट्रीय प्रौद्योगिकी संस्थान, राउरकेला

Dr. Bibekanand Mallick, **M.Tech., Ph.D.**
Assistant Professor

RNA Biology & Functional Genomics Lab.
Department of Life Science
National Institute of Technology
(*Ministry of H.R.D, Govt. Of India*)
Rourkela - 769 008, Odisha, India

Telephone: +91-661-246 2685 (O)
E-mails: vivek.iitian@gmail.com, mallickb@nitrkl.ac.in
Homepage: http://vvekslab.in

Date: 10. 05. 2013

CERTIFICATE

This is to certify that the thesis entitled "Identification of Molecular Signature in Epithelial carcinoma" submitted to National Institute of Technology; Rourkela for the partial fulfillment of the Master degree in Life science is a faithful record of bonafide and original research work carried out by Subhrata Jena under my supervision and guidance.

(Dr. Bibekanand Mallick)

ACKNOWLEDGEMENT

I wish to express my sincere thanks and gratitude to my guide Dr. Bibekanand Mallick, *Assistant Professor, Dept. of Life Science, National Institute of Technology, Rourkela, for his constant inspiration, encouragement and guidance throughout my project. I consider myself fortunate enough that he has given a decisive turn and boost to my career.*

I take this opportunity to express my indebtedness to my Professors for their enthusiastic help, valuable suggestions and constant encouragement throughout my work. I would also like to express my whole hearted gratitude to the Head of the department of life-sciences Dr. Samir Kumar Patra, and other faculty members, Dr. Surajit Das, Dr. Sujit Kumar Bhutia, Dr. Suman Jha, Dr. Bismita Nayak and Dr. Rasu Jayabalan for their good wishes, inspiration and unstinted support throughout my course.

I deeply acknowledge the constant support, encouragement, and invaluable guidance at every step of my project by, Devyani Samantarrai, Debashree Das Ph.D. scholar, Dept. of life science. I am obliged and thankful to them for providing me the opportunity to gain knowledge and understanding of working skills of the aspects of my work from them.

I take this opportunity to thank my friends Bini, Mitali, Bibhu for their throughout co-operation.

At the end, I bow down my head to the almighty whose omnipresence has always guided me and made me energized to carry out such a project.

Place: NIT, Rourkela *Subhrata Jena*

Date: 10th May, 2013

DEDICATED TO

MY

BROTHER

Dr. Abinash Rout

CONTENTS

LIST OF TABLES

Table No. **Page No.**

LIST OF FIGURES

Figure No. **Page No.**

ABSTRACT

ABSTRACT

Epithelial tumour or carcinoma is the most common cause of death in individual with cancer worldwide. Molecular basis of epithelial tumor is poorly understood. To elucidate the mechanism behind the abstractness of epithelial tumor, "Molecular Signature" is the more accurate and effective than possible standard approach. Microarray data analysis has made it possible to obtain molecular snapshot of genes of an organism at various disease state and experimental conditions. In this study, we discussed the uncovering of molecular signature from epithelial tumors (Brain, Stomach, Cervical cancer) on the basis of relative fold change and potential biomarker ability in cancer. To explore the molecular signature in epithelial tumor, we compared the gene expression profile of brain, stomach, cervical cancer. From microarray analysis we found 201 exclusive set of common genes in tumors of epithelial origin and from 201 genes, we are able to identify 10 genes that can be used as molecular signature for all types of cancer which has epithelial origin. Selected two genes (SERPINA3, SH3GL3) were experimentally validated by qRT-PCR in HeLa cell line. qRT-PCR established that these two genes are showing their up regulation with respect to Beta-Actin, which is a housekeeping gene. The identification of molecular signature has promising application for accurate detection, promote early diagnosis and screening of cancer.

INTRODUCTION

INTRODUCTION

Just as an individual's signature is unique and identifies him or her, the gene activity/expression profile of a cell or tissue can be a unique identifying molecular 'signature' for that cell or tissue. Molecular signatures or gene-expression signatures are one of the most significant translational developments of recent years. Since 1998, potential of molecular signatures has been established (www.nyuinformatics.org). Molecular signature is defined as *a set of biomolecular features (e.g. DNA sequence, DNA copy number, RNA, protein, and metabolite expression) together with a predefined computational procedure that applies those features to predict a phenotype of clinical interest on a previously unseen patient sample (Sung J et al., 2012).* Thus the unique pattern of gene expression for a given cell or tissue is referred to as its molecular signature.

As a structural and functional unit of organisms, normal cells are carefully programmed to participate in constructing the diverse tissues that make possible organism survival. When cells in a part of the body start to grow abnormally and uncontrollably, they are called tumour. There are several kinds of tumour. A tumor can be benign or malignant. The cells of benign tumours are close to normal in appearance and are not considered cancerous. They do not have the potential of being dangerous and grow slowly and do not invade nearby tissues or spread to other parts of the body. On the other hand a malignant tumour is cancerous and can spread beyond the original tumor to other parts of the body. They have a quite different and more focused agenda. They appear to be motivated by only one consideration: making more copies of themselves (Ames B. N, 1983). Cells, in an initially formed primary malignant tumour may travel to other parts of the body and seed new tumour colonies at distant sites in the body, where they begin to grow and form new tumours. This process is called metastasis. Angiogenesis is a key process in tumor growth and spread in which tumours provoke the growth of new blood vessels to the tumor from pre-existing vessels; these new blood vessels provide the tumor with oxygen and nutrients, allowing these cells to grow, invade nearby tissue and spread to other parts of the body .

Cancers are classified according to the embryonic origin of tissues and their histological structure. The international standard for the classification and nomenclature of histologies is the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3). From the histological point of view there are hundreds of different cancers, which are grouped into 4 major categories: epithelial tumours (carcinomas), connective tumours (sarcomas),

hematopoietic tissue tumours and nervous system tumours. Of these human, carcinomas, derived from epithelial tissues account for 80 to 90 percent of all cancer and sarcomas being relatively rare malignant tumours comprise less than 10% of all cancers (Jemal et al., 2003).

Carcinoma includes cancers of the breast, prostate, lung, pancreas and colon cancer. They have a number of subtypes which include adenocarcinoma, basal cell carcinoma, squamous cell carcinoma, and transitional cell carcinoma. Connective tissue tumors or sarcomas are further classified on the basis of the site of the tumor: bone or soft tissue (Lahat et al., 2008). Of these, soft tissue sarcoma (STS) is the collective term used for malignancies arising in muscles, fat, vessels and fibrous tissues (WHO classification). Nervous system cancer includes cancers of the different types of cells of the nervous system like glial cells, neuroepithelial cells, etc. Hematopoietic tissue tumours include leukemia, lymphoma and other related disorders. Here the cancer arise from hematopoietic (blood forming) cells that leave the bone marrow and tend to mature in lymph nodes or blood.

Gene expression signatures can be important both for diagnostic purpose and for providing information about the biological system underlying certain conditions by certain highlighting genes, which are both related to those conditions. Molecular signatures are often used to model patient's clinically relevant information like prognosis, survival time, etc, as a function of the gene expression data (Mehrabian et al., 2005; Hur et al., 2011). Individual genes used as a function of the gene expression data, are known as the signature components or metagenes. Each signature component show strong co-expression in different cancerous conditions. In cancer, use of signature is able to build a successful model, and able to predict the probability of developing stage of a metastasis.

Recent technological developments have made it possible to obtain high-resolution molecular snapshots of organisms, tissues, and even individual cells at various disease states and experimental conditions. Techniques in genomics, like DNA microarray are one of the most important development which are extremely useful for understanding gene functions and interactions. Gene expression profiling using microarray technology is one of the essential tools to monitor genome wide expression levels of genes in a given organism. Molecular signatures or gene-expression signatures are a key feature in many studies, and the use of microarray data for its generation is valuable in different disease diagnosis (Alizadeh et al. 2000; Pomeroy et al. 2002). Usually, microarray is a glass slide on to which DNA molecules

are fixed in an orderly manner at specific locations called spots. The most common tools used to carry out these measurements include complementary DNA microarrays (Schena, 1995), oligonucleotide microarrays (Lockhart et al. 1996) or serial analysis of gene expression (SAGE) (Velculescu et al. 1996). Microarrays may be used to measure gene expression in many ways, but one of the most common applications is to compare expression of a set of genes from a cell maintained in a particular disease condition to the same set of genes from a reference cell maintained under normal conditions. Significance of microarray data in basic research and target discovery is finding genes expressed in significantly different patterns across samples. It also includes biomarker determination to find genes that correlate with and presage disease progression. Generally microarray datas are available in Gene Expression Omnibus (GEO) Database of National Center for Biotechnology Information (NCBI). GEO contains large amount of array- and sequence-based data which are freely available for experimental as well as academic use. Many softwares are used to assists microarray data analysis like Gene spring, R, Bioconductor, dChip, Multi expression Viewer, MA Explorer, RELNET etc.

REVIEW OF LITERATURE

REVIEW OF LITERATURE

2.1. Cancer:

Cancer is the unregulated growth of cells resulting from gene mutations occurring due to various reasons. In cancer, cells divide and grow uncontrollably, forming malignant tumors, and invade nearby parts of the body. The cancer may also spread to more distant parts of the body through the lymphatic system or blood stream. Not all tumors are cancerous. Benign tumors do not grow uncontrollably, do not invade neighbouring tissues, and do not spread throughout the body. There are over 200 different known cancers that badly affect humans (Cancer Research UK, May, 2012). In 2008, an estimated 12.7 million new cancer cases were diagnosed worldwide. Lung, female breast, colorectal and stomach cancers were the most commonly diagnosed cancers, accounting for more than 40% of all cases. An estimated 7.6 million deaths from cancer occurred worldwide in 2008. Half of the cancer death is due to most common cancer like ling, stomach, liver, colorectal and female breast cancers (Cancer research, UK).

Common elements of staging of cancer are site of cancer, tumor size and number of tumors, lymph node involvement, cell type and tumor grade, the presence or absence of metastasis (National Cancer Institute). Cancer varies by stages. There are different stages of cancer. Hyperplasia, where the cells still appears normal, but is dividing too rapidly. The term Dysplasia used when cellular abnormality is restricted to the originating tissue. And in Invasive cancer, Cancer that has spread beyond the layer of tissue in which it developed and is growing into surrounding, healthy tissues. In Metastasis cancer, primarily formed malignant tumor may spread to other parts of the body and seed new tumor colonies. Invasive and metastasis cancer are looks like same, in both cases cells are invaded to the surrounding tissue and has potential to travel through the lymphatic system (Carlson et al., 2009). Metastasis is the principal event leading to death in individuals with cancer, yet its molecular basis is poorly understood (Sridher et al., 2002).

2.2. Classification of cancers:

According to the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3), tumors are classified based upon the morphology and topography of the neoplasm into 4 catagoies. These are epithelial cancer (carcinoma), connective tissue or mesenchymal cancer (sarcoma), Nervous system cancer, Hematopoietic cancer (leukemia and lymphoma).

Carcinoma: Carcinoma is one form of cancer that composed of cells, have developed the cytological appearance, histological architecture or molecular characteristics of epithelial cells (Berman, 2004). Epithelial cancer or carcinoma is most common type of cancer that occurring in humans. It begins in a tissue that lines the inner or outer surface of the body organ. Generally that has ectodermal or endodermal origin during embryogenesis. The majority of cancers in industrialized countries are solid tumors derived from epithelial tissues such as carcinomas of the breast, lung, gastrointestinal tract, and prostate (Woelfle et al., 2003). Usually, carcinomas have generally been classified in to various types using a combination of criteria including histology, Adenocarcinoma, Basal cell carcinoma, Squamous cell carcinoma, and Transitional cell carcinoma. Frequent origin site of carcinoma is lungs, breast, prostate, colon, pancreas, cervical, stomach and brain.

Brain Cancer (Glioblatoma): It is the most common and aggressive primary brain tumour that involves glial cells. Glioblastoma is a highly vascular tumor that expresses vascular endothelial growth factor, a key regulator of angiogenesis and tumor blood vessel permeability (Narita, 2013). The majority of cases (>90%) are primary glioblastomas that develop rapidly through de novo pathway, without clinical or histological evidence of a less malignant precursor lesion. Secondary glioblastomas develop through progression from low-grade diffusion of astrocytoma or anaplastic astrocytoma. During progression to glioblastoma, additional mutations accumulate, including loss of heterozygosity in 10q25-qter which is approximately 70% and it is the most frequent genetic alteration in both primary and secondary glioblastomas (Ohgaki and Kleihues, 2007).

Stomach (Gastric) Cancer: It is hard to diagnose stomach cancer in its early stages. Most stomach cancer is caused by Helicobacter pylori infection. Atrophy of acid-secreting parietal cells (PCs) frequently occurs during infection with Helicobacter pylori. During loss of hyaloruinic acid receptor, CD44 labelled undifferentiated cells in gastric unit of isthmus shows decrease proliferation of the gastric epithelium. CD44 binds with STAT3 and inhibition of either CD44 or STAT3 signalling causes decrease proliferation (Khurana et al., 2013).

Cervical cancer: Worldwide cervical cancer is the most common type of cancer in woman (Chepovetsky at al., 2013). It starts in the squamousal epithelium of cervix, which is the lower part of the uterus. All cervical cancers are caused by the infections of HPV (Human papilloma virus) appears to be a necessary factor in the development of almost all cases (90%) of cervical cancer (Walboomers et al., 1999). Infection is extremely common in young women. There are four major steps in cervical cancer development: infection of metaplastic epithelium at the

cervical transformation zone, viral persistence, progression of persistently infected epithelium to cervical pre-cancer, and invasion through the basement membrane of the epithelium (Schiffman et al., 2007).

Sarcoma: Connective tissue tumors or sarcomas are further classified on the basis of the site of the tumor: bone or soft tissue (Lahat et al., 2008). Of these, soft tissue sarcoma (STS) is the collective term used for malignancies arising in muscles, fat, vessels and fibrous tissues (WHO classification).Sarcoma developed from transformed cell of mesenchymal origin. Liposarcoma arises from fat where Leiomyosarcoma arises from smooth muscle.

Leiomyosarcoma: It is a rare malignant cancer of smooth muscle, accounts for 5-10% of cancers in smooth tissue muscle. It is resistant to chemotherapy and radiation and mainly found in any sites of the body like uterus (Arnold et al., 2012), stomach, small intestine, retroperitoneum (Piovanelloet al., 2007).

Liposarcoma: Liposarcoma is a malignant tumor that arises in fat cells of deep soft tissue and accounts for approximately 20% of all mesenchymal malignancies. They are large bulky tumor present inside the thigh or in the retroperitoneum.

Myxofibrosarcoma (MFS): Myxofibrosarcoma is one of the most common sarcomas in the extremities of elderly patients (Mentzel et al., 1996), present in trunk (12%), retroperitoneum or mediastinum (8%) and head (3%) (Mansoor and White, 2003). It characterised by distinct vascular patterns. Myxofibrosarcoma is a connective tissue neoplasm of fibroblastic origin set in a myxoid matrix and has been classified by some as a myxoid variant of malignant fibrous histiocytoma (MFH) (Wada et al., 2000; Mansoor and White, 2003). The majority of acral myxofibrosarcomas are observed in the leg rather than the arm, often attributed to the fact that the leg contains a greater volume of connective tissue and thus has a greater chance for malignant development (Hollowood et al., 1995; Ninfo et al., 1998; Arenson et al., 1986)

Hematopoietic cancer: Hematopoiesis gives rise to blood cells of different lineages throughout normal life. This cancer falls into two categories, Lymphoma and Leukemia. Abnormalities in this developmental program lead to blood cell diseases including lymphoma, when blood cancer occurs in B or T lymphocytes. Lymphocytes help to protect the body from various infection and disease. Abnormal lymphocytes divide faster than other normal cells. It develops in many parts of the body including the lymph nodes, spleen, blood and bone marrow (Medical News Today, 2013).

2.3. Microarrays Gene expression:

Microarrays Gene expression are prominent experimental tool in functional genomics allowing gaining of a deeper understanding of biological processes (Schena et al., 2005). In a single experiment, we can measure gene expression levels for thousands of genes and even an entire genome (Khalid et al., 2006).

Molecular diagnostics is a rapidly advancing field in which insights into disease mechanisms are being elucidated by use of new gene-based biomarkers. Some genes, proteins and some genetic variants are used to characterise the molecular signature of certain cancer. Microarray analysis helps to provide invaluable information on disease pathology, progression, resistance to treatment, ultimately leads to improved early diagnosis and innovative therapeutic approaches for cancer (Pascale et al., 2002).

2.4. Microarray Analysis tools:

R: R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS

Bioconductor: Bioconductor is an open source and open development software project analysis of genomic data, especially microarray data.

dChip: DNA- Chip Analyzer(dChip) is a model-based expression analysis of oligonucleotide arrays. It is a free software package, can be downloaded. It is also implemented in the Bioconductor package.

TM4: TM4 is a free suite of software tools for microarray analysis. It consists of 4 major applications: Microarray Data Manager (MADAM), Spotfinder, Microarray Data Analysis System (MIDAS) and Multiexperimant Viewer (MeV).

GenMAPP: Gene Map annotator and Pathway Profiler (GenMAPP) is free software for visualizing gene expression and proteomics data on pre-packaged and custom pathway (MAPPs).

GeneSpring: GeneSpring is widely-used commercial microarray analysis software.

Spotfire: Spotfire DecisionSite for Functional Genomics is commercial software for microarray analysis.

2.5. qRT-PCR: As we all know, the advent of Polymerase Chain Reaction (PCR) by Kary B. Mullis in the mid-1980s revolutionized molecular biology. At its most basic application, PCR can amplify a small amount of template DNA (or RNA) into large quantities in a few hours. This is performed by mixing the DNA with primers on either side of the DNA (forward and reverse), *Taq* polymerase (of the species *Thermus aquaticus*, a thermophile whose polymerase is able to withstand extremely high temperatures), free nucleotides (dNTPs for DNA, NTPs for RNA), and buffer. The temperature is then alternated between hot and cold to denature and reanneal the DNA, with the polymerase adding new complementary strands each time. In addition to the basic use of PCR, specially designed primers can be made to ligate two different pieces of DNA together or add a restriction site (Purves, et al. 2001). Recently,a new method of PCR quantification has been invented. This is called "real-time PCR" because it allows to viewing the increase in the amount of DNA as it is amplified. Several different types of real-time PCR are being marketed to the scientific community at this time, each with their advantages.

2.6. Cancer Cell lines: A HeLa cell is a cell type in an immortal cell line used in experimental work in research field. It is the oldest and most commonly used human cell line (Rahbari et al., 2009). The line was derived from cervical cancer cells taken on February 8, 1951(Scherer et al., 1951) from Henrietta Lacks. The cell line was found to be remarkably durable and prolific as illustrated by its contamination of many other cell lines used in research (Batts DW, 2010; Capes-Davis et al., 2010).

OBJECTIVES

OBJECTIVES

1. Microarray expression analysis of tumors of epithelial, mesenchymal and haematopoietic origin

2. Identification of overlapping sets of differentially expressed genes expressed in tumors of epithelial (Brain, Stomach, Cervical cancer), mesenchymal (Leiomyosarcoma, Liposarcoma and Myxofibrosarcoma) and haematopoietic origin tumours (lymphoma)

3. Uncovering the genes which are exclusively expressed in epithelial tumors (Brain, Stomach, Cervical cancer)

4. Establish a molecular signature profile for the epithelial tumors

5. Experimentally validate selected sets of genes that are abnormally differentially expressed

MATERIALS & METHODS

MATERIALS & METHODS

4.1. Gene Expression Data: Gene Expression is the process by which information of a given gene is used to synthesis a functional gene product (RNA or Protein). From the expression of a gene we can enumerate the concentration as well as the activity of the respective gene. Nonstandard amounts of gene product can be correlated with different diseases like cancer. Gene expression data are obtained from Gene expression omnibus (GEO) database of National Center for Biotechnology Information (NCBI). GEO is a public functional genomics data repository, which provides different tools to help the user query, access downloadable experimental data and curated gene expression profiles.

4.2. GEO data organisation:

4.2.1. GEO Series (GSE): A GEO series represents a curated collection of a group of samples corresponding to one publication, which are biologically and statistically comparable records. It may contain tables describing extracted data, summery conclusion or analysis. Each Series record is assigned a unique and stable GEO accession number.

4.2.2. GEO Sample (GSM): It is a sample record, which describes the conditions under which individual samples was handled. A sample entity must reference only one platform and may be include in multiple series. Each sample record is assigned a unique and stable GEO accession number.

4.2.3. GEO Platform (GPL): Stores the position and corresponding feature of each probe (spot) such as a GenBank accession number, open reading frame (ORF) name and clone identifier. Each Platform record is assigned a unique and stable GEO accession number. It is denoted as GPLxxx. A Platform may reference many Samples that have been submitted by multiple submitters. In our study, all these datasets are from one platform i.e. Affymetrix Human Genome U133A Array [HG-U133A], platform ID: GPL96.

For experimental performance, we used previously analysed datasets of different types of cancers with respect to normal tissue from GEO database. Three different types of cancers are taken for our analysis- carcinomas, sarcomas and haematopoietic cancers. In carcinomas, we considered Glioblastoma (Brain), Cervical and Stomach cancers; in sarcomas, liposarcoma, leiomyosarcoma, myxofibrosarcoma and in haematopoietic cancer, lymphoma for further molecular signature analysis. The gene expression or mRNA expression data

signifying two different criteria i.e. human normal tissues and cancerous tissues which we have considered for our study is given in Table 1.

4.3. Epithelial Cancers:

Table-1: The sources of Epithelial Cancers samples considered for analysis

Cancer Names	Source Name		Organism
	Test	Control/ Normal	
Glioblastoma(Brain)	Glioblastoma	Normal brain tissues	Homo sapiens
Cervical	Squamous cell Carcinoma	Normal cervix epithelium	Homo sapiens
Stomach	Primary gastric tumor	Normal gastric tissue.	Homo sapiens

4.4. Sarcoma:

Table-2: The source of sarcoma samples considered for analysis

Cancer Names	Source Name		Organism
	Test	Control/ Normal	
Leiomyosarcoma	Soft tissue sarcoma	Human control normal fat	Homo sapiens
Liposarcoma	Soft tissue sarcoma	Human Control normal fat	Homo sapiens
Myxofibrosarcoma	Soft tissue sarcoma	Human control Normal fat	Homo sapiens

4.5. Lymphoma:

Table-3: The source of Lymphoma samples considered for analysis

Cancer Names	Source Name		Organism
	Test	Control/ Normal	
Lymphoma	Burkitt's lymphoma	Lymphoblasts	Homo sapiens

Table-4: GEO accession numbers of the samples taken for molecular signature analysis

Cancer Names	GSEs		GSMs	
	Test	Control	Test	Control
Glioblatoma (Brain)	GSE8692	GSE2361	GSM215420 GSM215423 GSM215427	GSM339554 GSM339555 GSM339556
Cervical	GSE9750	GSE9750	GSM247650 GSM247651 GSM247652	GSM247188 GSM247189 GSM247190
Stomach	GSE15456	GSE2361	GSM44703 GSM44703 GSM44703	GSM387757 GSM387758 GSM387759
Leiomyosarcoma	GSE21122	GSE21122	GSM528322 GSM528323 GSM528324	GSM528425 GSM528426 GSM528427
Liposarcoma	GSE21122	GSE21122	GSM528402 GSM528403 GSM528404	GSM528428 GSM528429 GSM528430
Myxofibrosarcoma	GSE21122	GSE21122	GSM528348 GSM528349 GSM528351	GSM528431 GSM528432 GSM528433

Lymphoma	GSE1133	GSE1133	GSM18891 GSM18892	GSM18889 GSM18890

4.6. Microarray Analysis of Gene expression Data:

4.6.1. Retrieval of Gene expression Data:

- Datasets selection via the NCBI Entrez data retrieval system, which is keywords based.
- Triplicates: we took sample genes in triplicate to reduce the error occur during handling experiment, and to get the maximum correct result.
- Datasets selection from GEO database based on our requirement.
- GEO samples (GSM) were selected in triplicates; this was done to minimize the error rate.
- Raw data are provided as supplementary file for each GEO samples at the end of the section. Samples are downloaded from supplementary files in .CEL format.
- Raw files are downloaded as zip files. Files are unzipped and the raw file was extracted and renamed as “Test” and “Control” for easy handling while analysing in software.

4.6.2. Analysis of gene expression data:

GeneSpring GX is a tool specifically designed for biologists that need powerful statistical methods to analyze expression data. GeneSpring GX allows the user to understand the results of the statistical analyses in a biological context. GeneSpring GX provides the user with guidance through a complete analysis, providing results, fast. And it is an open-system platform.

- **Gene Spring GX user Interface:** The display pane contains several graphical views of dataset as well as algorithm results.
- The display pane is divided in to 3 parts:
  - The main GeneSpring GX Desktop in the center
  - The Project Navigation on the left
  - The GeneSpring GX workflow Browser and the Legend window in the right.

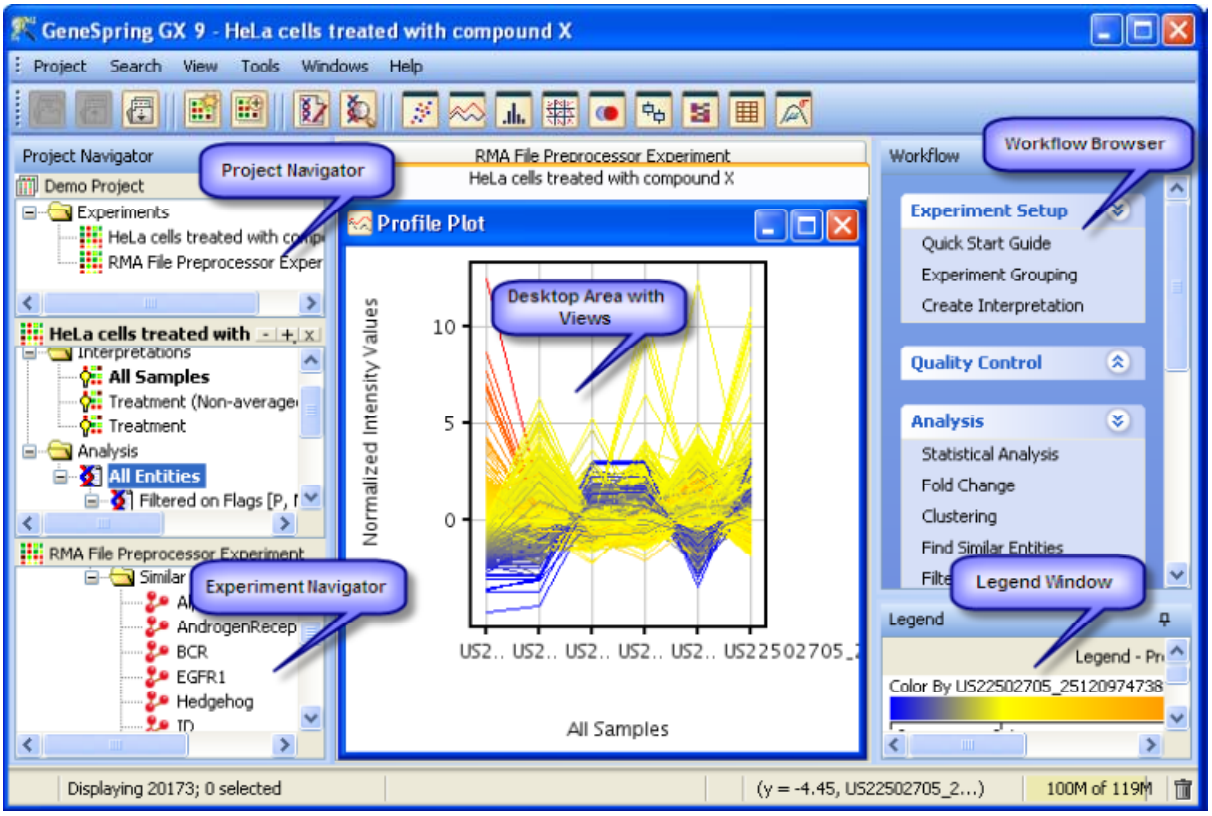

Figure 1: GeneSpring GX Layout

4.6.3. Organizational Elements in GeneSpring GX:

- Different experimental work in Gene Spring GX is organised in to Projects.
- A project comprises one or more related experiments.
- An experiment comprises Sample (Data source), interpretations (grouping of sample based on experimental parameters), analysis (Statistical steps ann associated results, typically entity lists).
- Each sample is associated with a chip type or its technology and will be imported and used along with a technology.
- An experiment comprises samples which all belong to the same technology.
- Here, we comprised the entire experimental sample under Affymatrix HG-U133A technology.
- Each sample has associated with experimental parameter.
- Each sample has to be given a parameter as Control and Test.

- **Experimental Grouping:** It requires to adding parameter to define the grouping and replicates structure of the experiment.
- **Significance Analysis:** Depending on the experimental groups, GeneSpring GX performs either T-test or ANOVA.

- Here, Sample grouping and significance based on the Normal and the Cancer sample with replicates was ANNOVA.

- **ANNOVA:** It is used for computing p-values, type of correction and P-value computation type.
  - Here, cut off P-value is less than 0.05.
- Multiple testing correction algorithm used was Benjamini-Hochberg FDR.

- **Fold Change Analysis:** Fold change analysis is used to identify genes with expression ratios or differences between a test and a control that are outside of a given cutoff or threshold. Fold change is calculated between any 2 conditions (Control Vs Test). The ratio between two conditions is calculated. It gives the absolute ratio of normalized intensities from the average intensities of the samples grouped. The entities satisfying the significance analysis are passed on for the fold change analysis.

  - Here, cut off value Fold change Value is more than 2.0
- **Retrieving of analyzed data:** After getting the result, we saved the files by clicking on 'save' button and for data retrieving; we right clicked on the generated file and then export the files onto the desktop.
- **Venn Diagram:** A Venn Diagram is used to show all the possible logical relation between a finite collection of gene entities. Generally, it reflects the union and intersection of entities passing the unpaired T-test values.
  - Here, in this experiment we got three Venn diagram using all the gene entities.
  - Epithelial cancers common genes Venn diagram (Brain cancer, Cervical cancer, Stomach cancer)
  - Mesenchymal cancers (Leiomyosarcoma, Liposarcoma, Myxofibrosarcoma) common genes Venn diagram.
  - Common genes between Mesenchymal, Epithelial and Hematopoietic origin cancers Venn diagram.
  - **Export the genes:** To export the genes exclusive to epithelial tissues, we selected the particular region by clicking on a particular arch of the Venn diagram and saved it first and then exported the file.
- **Heat Maps:** The Heat Map displays the normalized signal values of the conditions in the active interpretation for all the entities in the active entity list. The legend window

displays the interpretation on which the heat map was generated. The expression value of each gene is mapped to a colour-intensity value. The mapping of expression values to intensities depicted by a colour-bar created by the range of values in the conditions of the interpretation.

- Here, a concise Heat Map was generated by using extracted gene entities from Gene spring GX analysis.

- **Clustering:** Cluster analysis is a powerful way to organize genes or entities and conditions in the dataset into clusters based on the similarity of their expression profiles.
  - A variety of clustering algorithms: K-Means, Hierarchical, Self Organizing Maps (SOM), and Principal Components Analysis (PCA) clustering.
  - A variety of interactive views such as the ClusterSet View, the Dendrogram View, the Java Tree view and the U Matrix View are provided for visualization of clustering results.
  - **Hierarchical clustering:** Hierarchical clustering is one of the simplest and widely used clustering techniques for analysis of gene expression data. The method follows an agglomerative approach, where the most similar expression profiles are joined together to form a group. These are further joined in a tree structure, until all data forms a single group. The dendrogram is the most innate view of the results of this clustering method.
  - **Input parameters for clustering:** Clustering parameters are the entity list, the interpretation and the clustering algorithm. By default, the active entity list and the active interpretation of the experiment is selected.
  - Software used for Clustering: Cluster 3.0

4.7. Cluster 3.0: Cluster 3.0 is an enhanced version of cluster, which was originally developed by Michel Eisen while at Stanford University. The input file format is .txt file.Software used for visualisation of cluster result: JAVA Tree View

4.8. JAVA Tree View: Java Treeview is an open-source, cross-platform modify that handles very large datasets well, and with support of extensions file format that allow the results of additional analysis to be visualized and compared**.** The input file format is .cdt file.

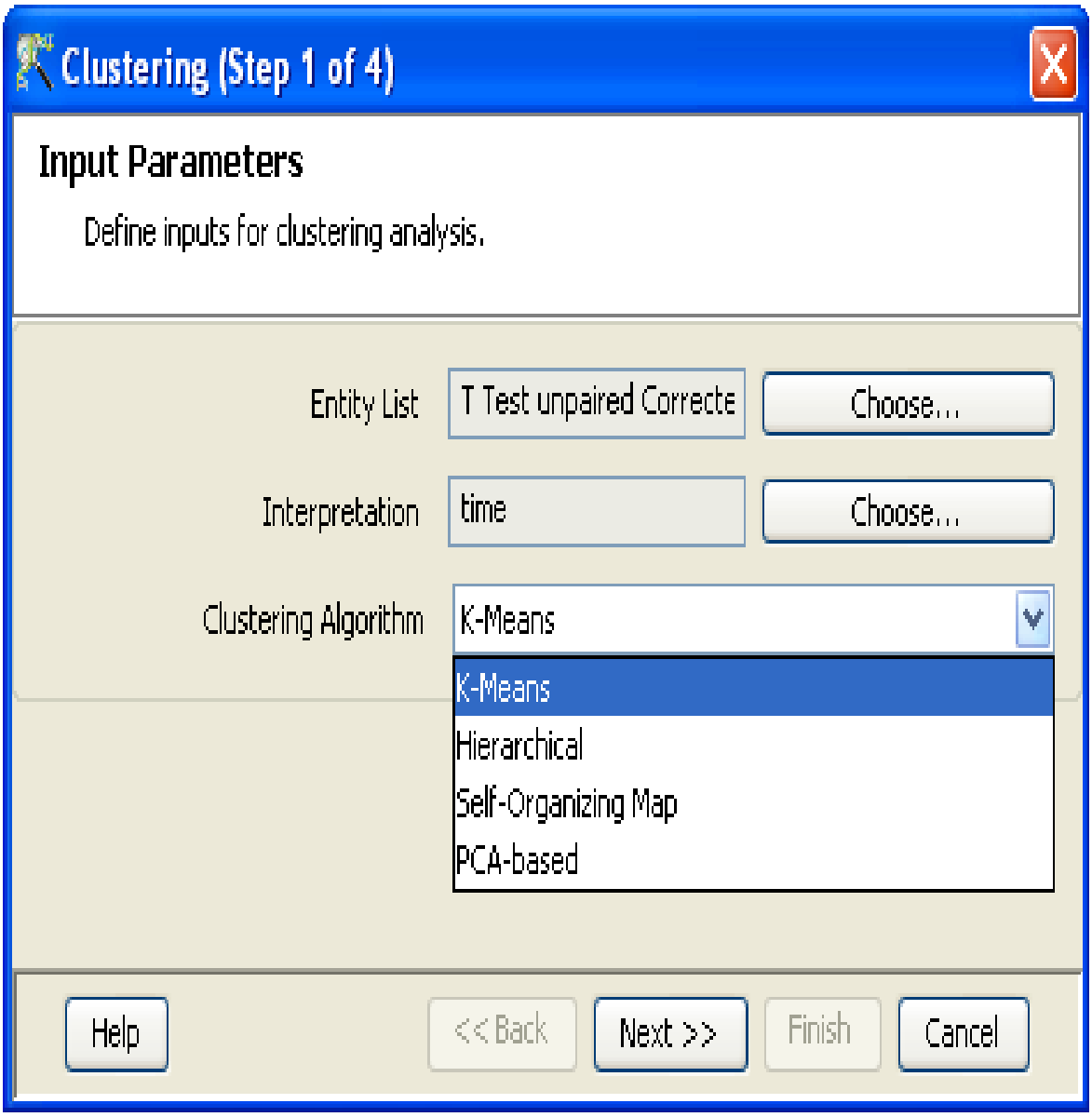

Figure 2: Clustering Wizard: Input parameters

4.8. Experimental Validation:

4.8.1. Cell line Culture:

- **Cell line:** These are the homogenous, fast growing cells which have desired properties. For experimental purpose cell lines were ordered from National Cancer Cell Science (NCCS), Pune, India.

- HeLa cell line was ordered. Because HeLa cells were the first human cells to grow well in the laboratory (Ewen Callaway, 2013)
- Cell lines are maintained in MEM medium with 10% FBS (HIMEDIA), 1% antibody solution (Penstrep solution FROM HIMEDIA).
- Subculture of cell lines: Subculture of cells required to relax the load of cells in the medium. The colour changes from red to orange then pale yellow indicates the sufficient utilization of media by cells.
- Followed Protocol:
    - Remove the used media.
    - Washed the cell lines with PBS buffer, to take out the FBS supplement.
    - Trypsin EDTA was added to detach the adherent cells from the flask surface.
    - Equal amount of MEM media was added to neutralize the trypsin.
    - Then the cells were centrifuged at 1000 rpm for 5 minutes.
    - Trypsin and media was decanted.
    - 5 ml of MEM was added to cell again and mixed properly by tapping and cells were counted in a haemocytometer.
    - Finally, cell with fresh medium was incubated in CO_2 incubator where level of CO_2 maintained is 5%.
    - After 48-72 hours, cells are fully grown and ready for RNA isolation.

4.8.2. RNA Isolation: RNA isolation is carried out using the cultured cell lines by the help of Qiagen RNA isolation kit (RNeasy Kit).

Followed protocol:

1. A maximum of 1×10^7 cells were harvested as a cell pellet and the the appropriate volume of Buffer RLT (Lysis buffer) was added.
2. 1 volume of 70% ethanol was added to the lysates and mixed well by pipetting. It should not be centrifuge.
3. 700 µl of the sample, including any precipitation was transfered to an RNeasy Mini spin column placed in a 2ml collection tube (supplied). It was centrifuged for 15s at ≥8000×g (13,000 RPM). Flow –through was discarded.
4. Then 700 µl Buffer RW1 (wash buffer) was added to the RNeasy spin column. It was centrifuged for 15s at ≥8000×g (13,000 rpm). Flow –through was discarded.

5. 500 µl Buffer RPE (wash buffer) was then added to the RNeasy spin column. It was centrifuged for 15s at ≥8000×g (13,000 rpm). Flow –through was discarded.
6. 500 µl Buffer RPE was again added to the RNeasy spin column. It was centrifuged for 15s at ≥8000×g (13,000 rpm). Flow –through was discarded.
7. The RNeasy spin column was then placed in the new 1.5 ml collection tube and 30-50 µl of RNase- free water was added directly to the spin column membrane. It was centrifuged for 15s at ≥8000×g (13,000 rpm) to elude the RNA**.**
8. Using Nanodrop (Eppendorf) the purity and quantity of RNA yield is checked by taking only 1-2µl of the sample.

4.8.3. cDNA Synthesis:

cDNA synthesis was carried out using SuperScript First-Strand Synthesis System for RT-PCR by Invitrogen using oligo dT primers.

The steps in cDNA synthesis:

1. Each component of the kit was mixed and briefly centrifuge before use.
2. For each reaction, the following components were combined in a sterile 0.2 or 0.5ml tube.

Components	Amount
RNA (2µg)	5 µl
10 mM dNTP mix	1 µl
Primer (0.5µg/µl oligo $(dT)_{12-18}$)	1µl
DEPC treated water	10µl

3. The RNA/primer mixture was incubated at 65˚c for 5 minutes and then placed on ice for at least 1 minute.
4. In a separation tube, the following 2X reaction mix was prepared, by adding each component in the indicated order.

Components	1RXn	10 RXnS
10X RT buffer	2 µl	20 µl
25mM $Mgcl_2$	4 µl	40 µl
0.1M DTT	2 µl	20 µl
RNase out ^{TM} (400/ µl)	1 µl	10µl

5. 9µl of the 2X reaction mixture was added to each RNA/primer mixture from step 3 and mixed gently and collect by briefly centrifugation.
6. It was incubated at 42·c for 2 minutes.
7. 1µl of super script™ II RT was then added to each tube.
8. Incubated at 42·c for 50 minutes.
9. Then, terminated the reaction at 70·c for 15 minutes. Chilled on ice.
10. The reaction was collected by brief centrifugation and 1µl of RNase H was added to each tube and incubated for 20 minutes at 37·c.
11. Then, the reaction is stored at -20·c or used for PCR immediately.

4.8.3. qRT PCR:

This protocol describes the detailed experimental procedure for real-time RT-PCR using SYBR Green. The procedure begins with reverse transcription of total RNA. The cDNA is then used as template for real-time PCR with gene specific primers.

Materials required:

- Oligonucleotide Primers: Gene specific primers are retrieved from primer bank. These primers are ordered from the SIGMA Genosys.
- cDNA.
- SYBR Green PCR master mix.
- White optical tube and cap strips.
- RT-PCR (Eppendorf).

1. **PRIMERS:**

   Control gene: β- Actin (Housekeeping Gene)

   Test genes: SERPINA3, SH3GL3

2. **Primer Concentration:** (Forward & Reverse)

   Stock: 100µM

   Working Concentration: 10µM

   Primer Concentration in experiment: 500nM

   For 2 genes (3 replicates each): 70µl solution required including primers.

   Forward Primer: 3.5 µl

   Reverse Primer: 3.5 µl

3. **Fluorescent probe:** SYBR Green mix- SYBR green, dNTP, Mg^{2+}

   SYBR Green concentration in experiment: 35µl (1X)

4. **cDNA diluted to 1:20 ratio:** 40 µl prepared.

   Required cDNA: 28 µl

5. **Oligonucleotide Primers**:

Table-5: Primer sequence of Control and Test genes

Oligo Name	5'<-----Sequence--->3'	Lenght
SERPINA3 F	GATCGGGCATCACCTGAAAAA	21
SERPINA3 R	TCGTCTGGTATCTTACCTACTCG	23
SH3GL3 F	CCTGAAGGCCCCTGATAAGAA	21
SH3GL3 R	GCTGGACTGATTGAGGGTGC	20
β-Actin F	CATGTACGTTGCTATCCAGGC	21
β-Actin R	CTCCTTAATGTCACGCACGAT	21

Table-6: Cycles temperature in qRT PCR

STAGE	TEMPERATURE (°C)	TIME	CYCLE
Stage 1	95	20 sec	1
Stage 2	95 55 68	15 sec 15 sec 20 sec	40
Stage 3	95 60 95	15 sec 15 sec 15 sec	1

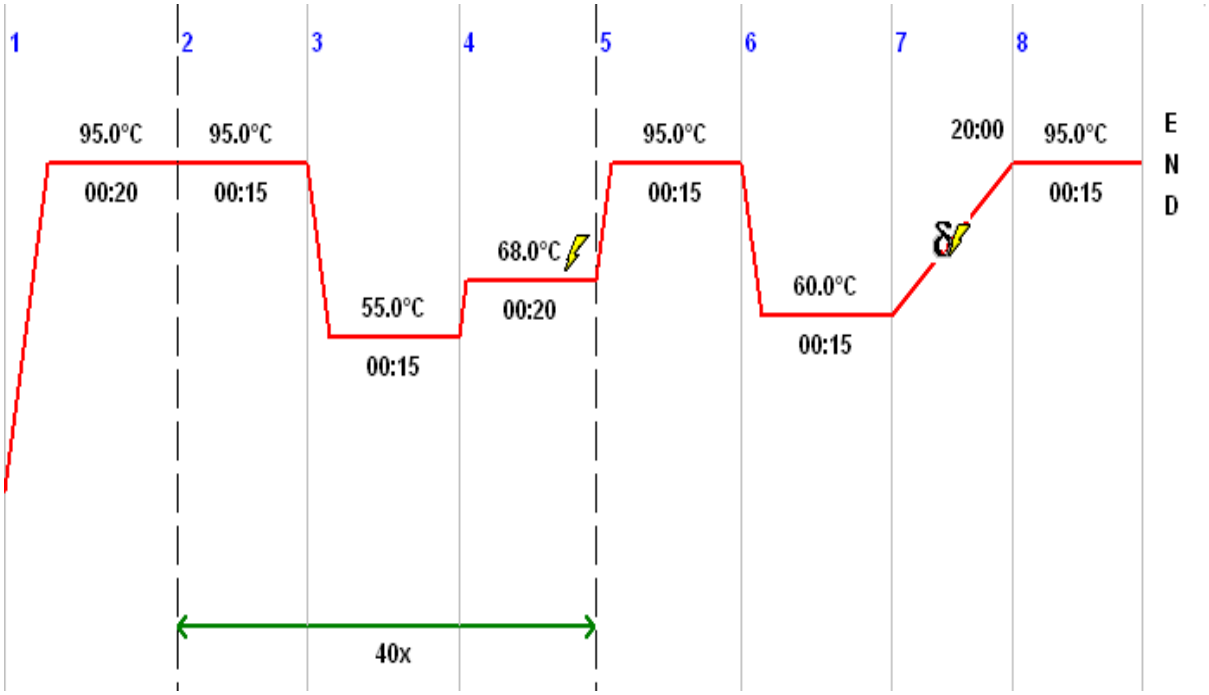

Figure 3. Cycle temperature and time for qRT-PCR

RESULTS AND DISCUSSIONS

RESULTS & DISCUSSIONS

5.1. Microarray analysis:

- From microarray analysis using GeneSpring GX, we obtained 348 differentially expressed genes expressed common to cancers of epithelial origin, i.e. brain, stomach, cervical cancer.

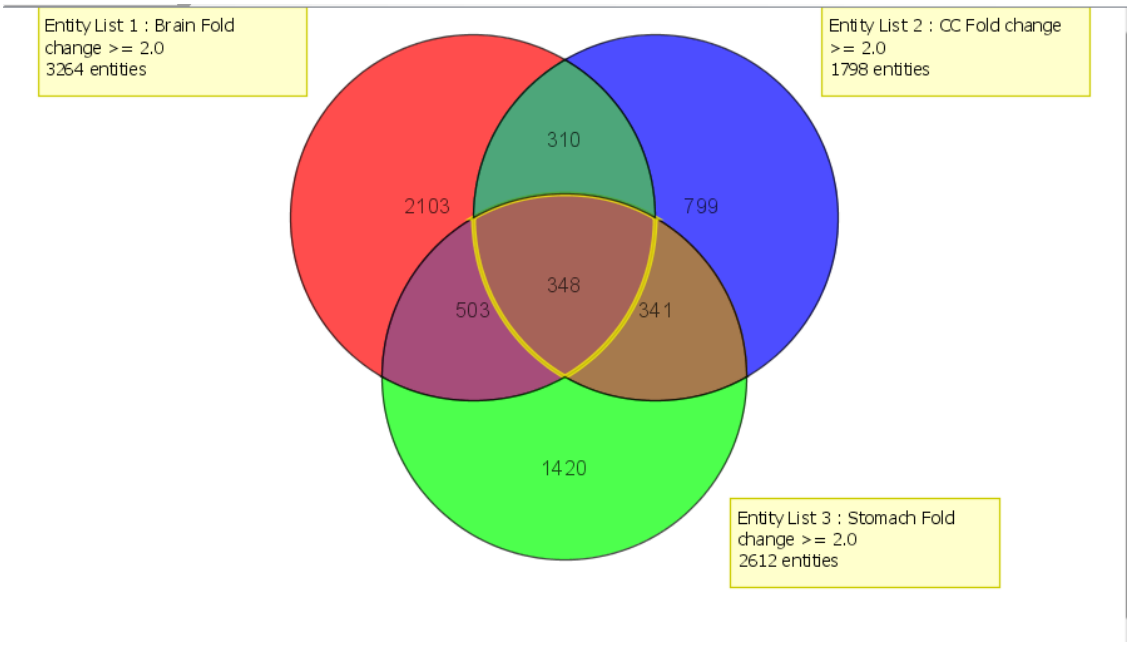

Figure 4. Venn diagram representing the common genes expressed in epithelial origin cancer (Brain, Stomach, Cervical cancer)

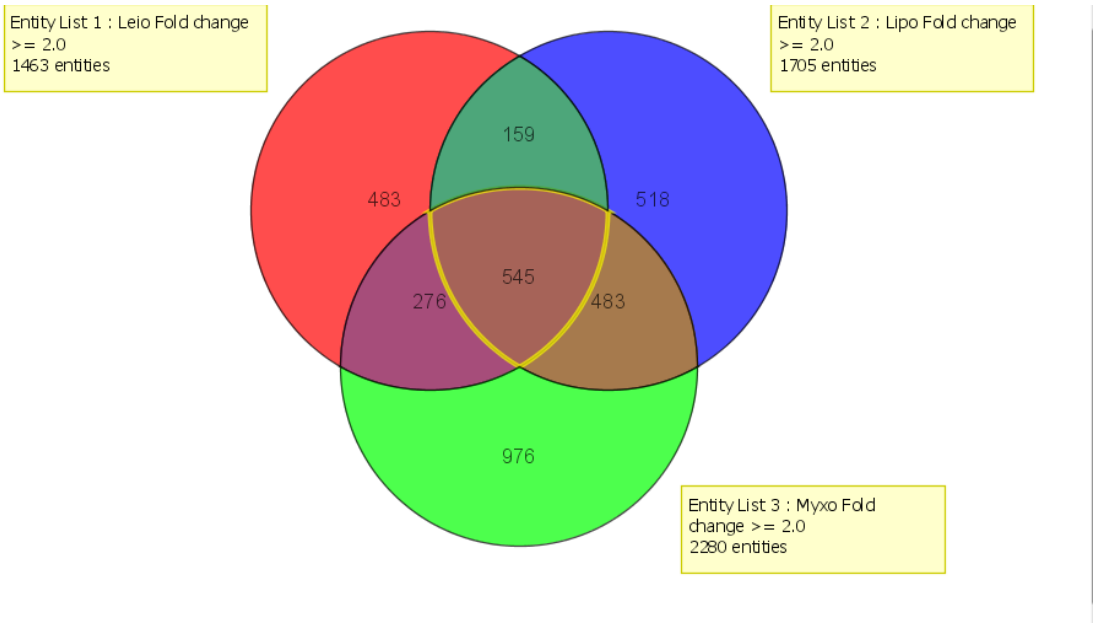

Figure 5. Venn diagram representing the common genes expressed in mesenchymal origin cancer

- From mesenchymal tumor (Leiomyosarcoma, Liposarcoma and Myxofibrosarcoma) , we obtained 545 common genes, which has mesenchymal origin.

- By comparing common genes differentially expressed in all three types of cancer of different origin (epithelial, mesenchymal and lymphoma), we got 201 genes that are exclusively differentially expressed in cancers of epithelial origin.

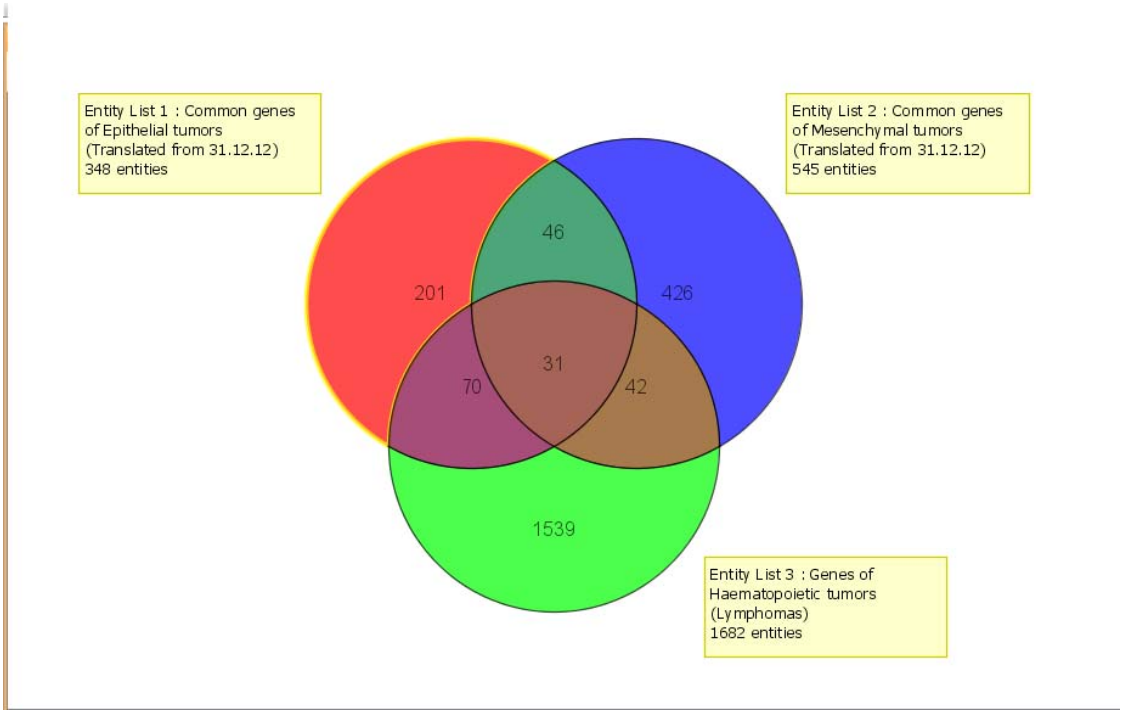

Figure 6 : Venn diagram representing the common genes & exclusive sets of genes between Epithelial, mesenchymal and Lymphoma origin cancer

5.2. Clustering:

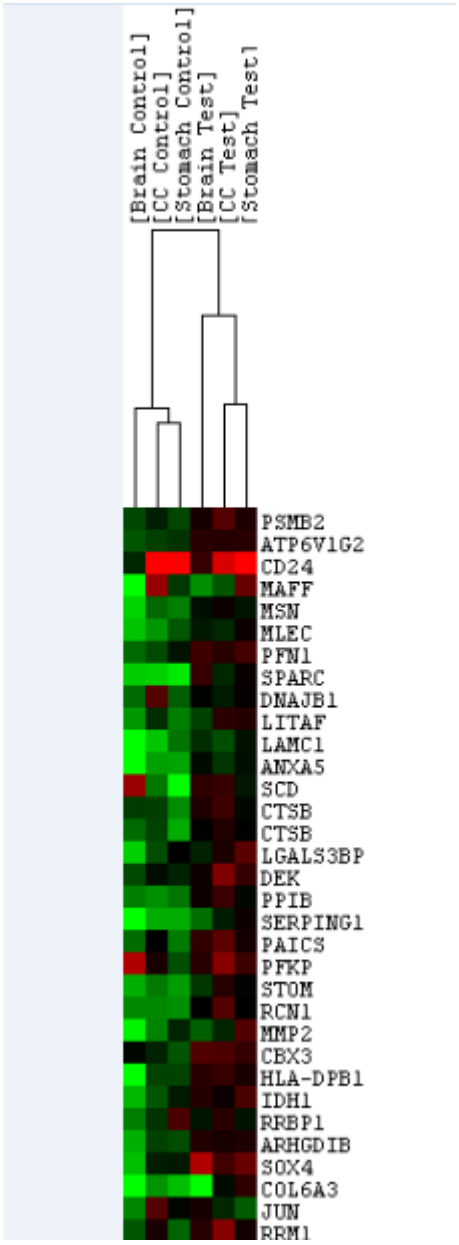

Figure 7 : Heat Map of microarray data of epithelial origin cancers

5.3. Gene list analysis:

- According to our objectives, we selected exclusive epithelial origin cancer related genes for further molecular signature analysis. The exclusive sets of epithelial origin cancer contained 201 numbers of genes, which are exclusively present in Brain, Cervical and stomach cancer.

Table 7: Exclusive set of genes in epithelial origin tumors (Stomach, Cervical, Brain cancer)

Gene Symbol	Regulation in Stomach cancer	Fold change in stomach cancer	Regulation in cervical cancer	Fold change in cervical cancer	Regulation in brain cancer	Fold change in brain cancer
PSMB2	Down	2.2474647	down	2.591263	Down	2.153261
ATP6V1G2	Down	2.1245968	down	2.277777	Down	2.827335
CD24	Down	2.1388943	up	2.543346	Down	2.133525
MAFF	Down	3.7758126	up	6.990386	Down	3.791928
MSN	Down	2.3790982	down	2.603572	Down	4.95187
MLEC	Down	2.2913814	down	2.465684	Down	4.08609
PFN1	Down	2.0546515	down	2.540145	Down	4.025861
SPARC	Down	7.3933825	down	4.089357	Down	8.2469
DNAJB1	Down	2.2949274	up	2.43908	Down	2.361121
LITAF	Down	3.858532	down	2.070647	Down	2.053864
LAMC1	Down	2.2272682	down	2.6912	Down	11.05916
ANXA5	Down	3.4218462	down	2.419475	Down	7.230515
SCD	Down	7.347266	down	4.224112	Up	2.400863
CTSB	Down	3.0131385	down	2.760871	Down	2.126846
CTSB	Down	4.0379577	down	2.283448	Down	2.46261
LGALS3BP	Down	2.1090496	down	2.353593	Down	4.16987
DEK	Down	2.0504317	down	3.16221	Down	2.067486
PPIB	Down	2.4568062	down	5.211479	Down	3.167322
SERPING1	Down	4.608875	down	3.298965	Down	3.461534
PAICS	Down	3.2486641	down	2.143968	Down	3.642463
PFKP	Down	3.1139922	down	2.45083	Up	3.131392
STOM	Down	3.5062802	down	3.729752	Down	2.725554
RCN1	Down	3.240057	down	6.09659	Down	3.128266
MMP2	Down	2.6291802	down	2.191986	Down	3.399188
CBX3	Down	3.124692	down	2.49818	Down	2.044873
HLA-DPB1	Down	2.130417	down	2.721022	Down	15.79708
IDH1	Down	2.2729692	down	2.245146	Down	6.186771
RRBP1	Up	2.1497543	down	2.12737	Down	2.376985
ARHGDIB	Down	2.4134653	down	2.098466	Down	5.663301
SOX4	Down	2.8947997	down	2.18217	Down	20.64291
COL6A3	Down	7.0182147	down	3.064116	Down	2.788027
JUN	Up	2.0348654	up	2.773805	Down	3.680503
RRM1	Down	2.6544075	down	2.407301	Down	2.748774
G3BP1	Down	3.7455091	down	2.078785	Down	2.401074
LAMB1	Down	3.338808	down	2.26632	Down	3.452669
TGFBI	down	3.7333894	down	3.738732	Down	8.903104
MCM3	down	2.7512023	down	3.999191	Down	3.881331
IFITM1	down	3.9529696	down	2.662751	Down	2.002034
IER3	down	2.0629313	up	3.947531	Down	2.324432
UBE2L6	down	2.033642	down	2.695916	Down	2.32301
TIMP1	down	5.8077855	down	2.526132	Down	14.71475
LAPTM5	down	2.589336	down	4.104724	Down	8.402963
SGK1	up	2.7926662	up	2.311896	Down	2.225214

CD14	down	2.0339742	down	4.395381	Down	18.04268
LUM	down	2.8608117	down	4.953169	Down	4.807593
MTHFD2	down	2.0601537	down	7.606896	Down	10.24751
AEBP1	down	3.5108612	down	3.393021	Down	2.498676
LBR	down	2.3810449	down	2.657699	Down	6.953003
LPCAT1	down	2.7648711	down	4.416288	Down	2.879076

- From 201 genes, we shortlisted 10 genes on the basis of standard fold change value (10 fold change) and biomarker potential in different epithelial carcinoma.
- From shortlisted 10 genes, two genes are taken for experimental validation.
- Selected genes: SERPINA3, SH3GL3

Table 8: Genes shortlisted for further analysis & validation

Gene names	Regulation
CD14	Down
COLIA1	Down
SOX4	Down
MVP	Down
CTSL2	Down
SERPINA3	Down
GULP	Up
NEBL	Up
SH3GL3	Up
FXYD	Up

SERPINA3 (NCBI, Gene ID- 12**)**:

- SERPINA3 has different aliases, Serpin peptidase inhibitor, clade4 (alpha1antiproteinase, antitrypsin member3) or Alpha-1 antichymotrypsin. Serine or cystein proteinase inhibitor, clade4, mem-3. It's a member of the serine protease inhibitor class. It comes under cell-growth inhibiting gene. It has plasma protein inhibiting capacity. Polymorphism in this gene leads to protease targeting. It is tissue specific. Variation causes Alzheimer's disease and deficiency causes liver disease. Mutation in SERPINA3 causes Parkinson's disease & chronic pulmonary disease. Alpha-1 anticymotrypsin is an acute phase protein, which level increases in acute &

chronic inflammation. It is related to JAK STAT pathways. The A/T polymorphism in SERPINA3 gene influences expression of this protein, while T allele was strongly down regulated; it induced the high level of cathepsin activit.

SH3GL3 (NCBI,Gene ID-6457)**:**

- SH3GL3 has different aliases domain GRB2-like protein 3.This gene is implicated in endocytosis. Present in cytoplasm as a peripheral membrane protein. Protein of SH3GL3 is endophilin-A3 contains 347 amino acids, weight is 39285 Da. No specific pathways.

5.4 Experimental validation:

RNA isolation:

- **260/280 Ratio:** This ratio indicates the absorbance of DNA and RNA at 260 nm and 280 nm, which is used to assess the purity of DNA and RNA. A ratio is expected approximately 1.8 and generally accepted as "pure" for DNA; a ratio of approximately 2.0 is generally accepted as "pure" for RNA. If the ratio is significantly lower in either case, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm.
- **260/230 Ratio**: This ratio is used as a secondary measure of nucleic acid purity. The 260/230 values for "pure" nucleic acid are often higher than the respective 260/280 values. Expected 260/230 values are normally in the range of 2.0-2.2. If the ratio is significantly lower than expected, it may indicate the presence of contaminants which absorb at 230 nm.
- Here, we got following results of two samples; Sample1 has perfect ratio absorbance in 260nm and 280 nm wavelengths where Sample2 has less value.

- **Sample1** = 500.6 µg/ml
  At (260/280) ratio = 1.99
  At (260/230) ratio = 2.01

- **Sample 2** = 125.2 µg/ml
  At (260/280) ratio = 1.54
  At (260/230) ratio = 1.10

qRT-PCR :

- Generally qRT-PCR melting curve analysis used to quantify nucleic acid, mutation detection and for genotype analysis.
- From qRT PCR analysis of SERPINA3, we got fine melting temperature curve of SERPINA3 in comparison with control gene, β-actin. Melting temperature of SERPINA3 is 80·c. Samples of gene taken in triplicates. Three picks of SERPINA3 positioned at one place. Relative quantification with respect to control gene representing, high expression of SERPINA3.

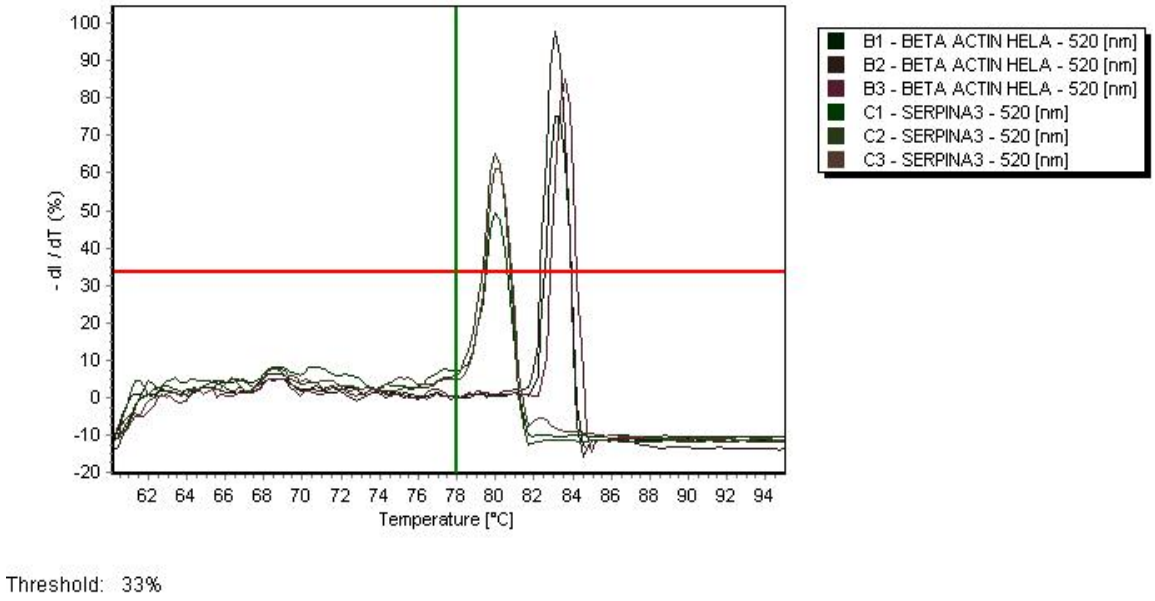

Figure 8: Relative quantification result of SERPINA3 with respect to control gene, β- Actin in qRT PCR analysis

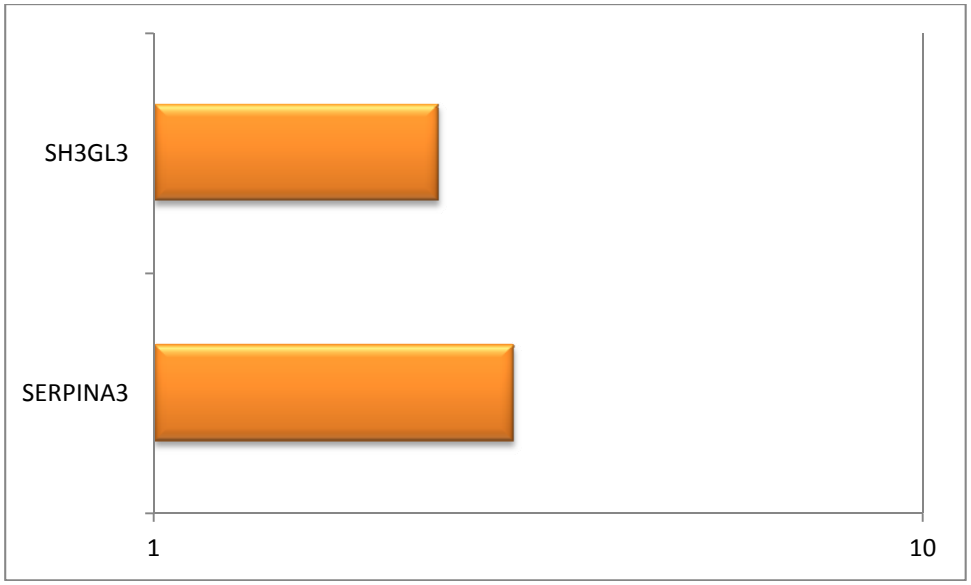

Figure 9: Relative expression of SERPINA3 and SH3GL3 with respect to control

- From the relative gene expression study of SERPINA3 and SH3GL3, it is cleared that both are up regulated in Hela cell line.

CONCLUSION

CONCLUSION

By studying the gene expression of brain, stomach, cervical cancer; we are able to identify a set of genes that are exclusively expressed in cancers of epithelial origin that can be used as molecular signature for all types of cancers that originated in epithelial tissues. By using these molecular gene expression signatures, we will be able to diagnose tumors for their origin where EMT has occurred. Further screening of 201 exclusive set of genes gave rise to 10 genes which are proposed to play a significant role in cancers of epithelial origin, out of which 2 genes (SERPINA3, SH3GL3) were experimentally validated in HeLA cancer cell lines by RT-PCR method for their expression.

qRT-PCR established that these two genes are showing their up regulation in comparison to control gene, β-actin. β-actin is a housekeeping gene, it expressed in all cells of an organism under normal and patho-physiological conditions. Even though in Microarray data analysis SERPINA3 to be down regulated (Kloth et. al., 2008) and SH3GLE3 has up regulated expression (Fang et. al., 2012 and Nguyen et. al., 2007), but qRT PCR showed up regulation which gave the idea that there might be some error in microarray data analysis or during validation. To know its regulation further well in epithelial tumors we will have to check its expression in normal cell lines too.

Several interesting candidate genes can have potential as therapeutic targets that are obtained from our present analysis. Validation of gene expression signatures in larger series needs to be performed to improve the understanding of the metastatic process of epithelial cancer further. SERPINA3 is a serine protease inhibitor (Serpins) which belongs to the Serpins superfamily with inhibitory activity. SERPINA3 has versatile role and in associated with inflammatory reaction in malignant melanoma and gastric cancer. It's a potential biomarker in colorectal cancer (CRC) progression (Kloth et. al., 2008). We considered that the up regulation expression of SERPINA3 may be considered as signature in epithelial carcinoma like brain, stomach, cervical cancer. Furthermore, the second gene SH3GL3 is associated with colorectal cancer and is a potential biomarker (Fang et. al., 2012 and Nguyen et. al., 2007) and same can be assumed for cancers of epithelial origin, but needs further validation. For establishing our gene-expression signature that distinguishes epithelial cancer from mesenchymal as well as hematopoietic cancer with high confidence, further validation of many differentially expressed genes are needed.

REFERENCES

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** Nature. 2000 Feb 3;403(6769):503-11.
2. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR. **Prediction of central nervous system embryonal tumour outcome based on gene expression**. Nature. 2002 Jan 24;415(6870):436-42.
3. American Cancer Society. **Cancer Facts and Figures** *2012.* Atlanta, Ga: American Cancer Society; 2012
4. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O.**Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 270, 467–470(1995).
5. Velculescu**, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W.** Serial analysis of gene expression. *Science* 270, 484–487 (1995).
6. Lockhart, D. J. *et al.* **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nature Biotechnol.* 14, 1675–1680 (1996).
7. Kloth JN, Gorter A, Fleuren GJ, Oosting J, Uljee S, ter Haar N, Dreef EJ, Kenter GG, Jordanova ES. **Elevated expression of SerpinA1 and SerpinA3 in HLA-positive cervical carcinoma**. J Pathol. 2008 Jul;215(3):222-30. doi: 10.1002/path.2347. PubMed PMID: 18438953.
8. Fang WJ, Zheng Y, Wu LM, Ke QH, Shen H, Yuan Y, Zheng SS. **Genome-wide analysis of aberrant DNA methylation for identification of potential biomarkers in colorectal cancer patients.** Asian Pac J Cancer Prev. 2012;13(5):1917-21. PubMed, PMID: 22901147
9. Nguyen ST, Hasegawa S, Tsuda H, Tomioka H, Ushijima M, Noda M, Omura K, Miki Y. **Identification of a predictive gene expression signature of cervical lymph node metastasis in oral squamous cell carcinoma**. Cancer Sci. 2007 May;98(5):740-6. Epub 2007 Mar 28. PubMed PMID: 17391312.

10. Ewen Callaway. **Most popular human cell in science gets sequenced. 2013 march 15, Nature publication**.
11. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, 270:467-470
12. Carlson, R. W.; Allred, D. C.; Anderson, B. O.; Burstein, H. J.; Carter, W. B.; Edge, S. B.; Erban, J. K.; Farrar, W. B. Et al. (2009). "**Breast cancer. Clinical practice guidelines in oncology**". *Journal of the National Comprehensive Cancer Network : JNCCN* 7 (2): 122–192.
13. Fang WJ, Zheng Y, Wu LM, Ke QH, Shen H, Yuan Y, Zheng SS. **Genome-wide analysis of aberrant DNA methylation for identification of potential biomarkers in colorectal cancer patients**. Asian Pac J Cancer Prev. 2012;13(5):1917-21.
14. Woelfle U, Cloos J, Sauter G, Riethdorf L, Jänicke F, van Diest P, Brakenhoff R, Pantel K. **Molecular signature associated with bone marrow micrometastasis in human breast cancer**. Cancer Res. 2003 Sep 15;63(18):5679-84
15. Berman JJ. Tumor **taxonomy for the developmental lineage classification of neoplasms.** BMC Cancer. 2004 Nov 30;4:88. PubMed PMID: 15571625; PubMed Central PMCID: PMC535937
16. Narita Y. Drug Review: **Safety and Efficacy of Bevacizumab for Glioblastoma and Other Brain Tumors.** Jpn J Clin Oncol. 2013 Apr 12. [Epub ahead of print] PubMed. PMID: 23585688.
17. Khurana SS, Riehl TE, Moore BD, Fassan M, Rugge M, Romero-Gallo J, Noto J,Peek RM, Stenson WF, Mills JC. **The hyaluronic acid receptor CD44 coordinates normal and metaplastic gastric epithelial progenitor cell proliferation**. J BiolChem. 2013 Apr 15.
18. Chepovetsky J, Kalir T, Weiderpass E., **Clinical applicability of microarray technology in the diagnosis, prognostic stratification, treatment and clinical surveillance of cervical adenocarcinoma**. Curr Pharm Des. 2013;19(8):1425-9.PubMed PMID: 23016775.
19. Crosbie EJ, Einstein MH, Franceschi S, Kitchener HC. **Human papillomavirus and cervical cancer.** Lancet. 2013 Apr 22. doi:pii: S0140-6736(13)60022-7.10.1016/S0140-6736(13)60022-7.
20. Medical News Today."**What is lymphoma".** Retrieved 28 February 2013

21. Piovanello P, Viola V, Costa G, Carletti M, Cecera A, Turchetta F, Iudicone R,Catalano G, Santucci A, Recchia F, Fiorillo L, Menichella MA, Baiano G. Locally Advanced **leiomyosarcoma of the spleen: A case report and review of the literature**. World J Surg Oncol. 2007 Nov 28;5:135. PubMed PMID: 18045454; PubMed Central PMCID: PMC2221972.
22. Arnold LM 3rd, Burman SD, O-Yurvati AH**. Diagnosis and management of primary pulmonary leiomyosarcoma**. J Am Osteopath Assoc. 2010 Apr; 110(4):244-6. PubMed PMID: 20430913
23. Mentzel T, Calonje E, Wadden C, Camplejohn RS, Beham A, Smith MA, Fletcher CD. **Myxofibrosarcoma. Clinicopathologic analysis of 75 cases with emphasis on the low-grade variant.** Am J Surg Pathol. 1996 Apr;20(4):391-405. PubMed PMID:8604805.
24. Mansoor A, White CR. **Myxofibrosarcoma presenting in the skin: clinicopathological features and differentialdiagnosis with cutaneous myxoid neoplasms**. Am J Dermatopathol 2003; 25: 281-286
25. Hollowood K, Fletcher CD. **Soft tissue sarcomas that mimic benign lesions**. Semin Diagn Pathol 1995
26. Ninfo V, Montesco MC. **Myxoid tumors of soft tissues: a challenging pathological diagnosis.** Adv Clin Path 1998.
27. Arenson DJ, Miceli JS, Bush WJ and Hussain A. **Myxofibrosarcoma of the lower extremity. A case report**. J Am Podiatr Med Assoc 1986; 76:102-105
28. Schena M, Shalon D, Davis RW, Brown PO. **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. Science. 1995 Oct 20;270(5235):467-70
29. Sabah Khalid, Karl Fraser, Mohsin Khan, Ping Wang, Xiaohui Liu, Suling Li, 2006, **Analysing Microarray Data using the Multi-functional Immune Ontologiser, Journal of Integrative Bioinformatics.**
30. Pascale F. Macgregor1 and Jeremy A. Squir, **Application of Microarrays to the Analysis of Gene Expression in Cancer**, 2002, Cancer Diagnostic Review.
31. Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, Snijders PJ, Peto J, Meijer CJ, Muñoz N. **Human papillomavirus is a necessary cause of invasive cervical cancer worldwide**. J Pathol. 1999.

32. Rahbari R, Sheahan T, Modes V, Collier P, Macfarlane C, Badge RM. **A novel L1 retrotransposon marker for HeLa cell line identification. Biotechniques**. 2009 Apr;46(4):277-84. doi: 10.2144/000113089.
33. Capes-Davis A, Theodosopoulos G, Atkin I, Drexler HG, Kohara A, MacLeod RA,Masters JR, Nakamura Y, Reid YA, Reddel RR, Freshney RI. **Check your cultures! A list of cross-contaminated or misidentified cell lines.** Int J Cancer. 2010 Jul1;127(1):1-8. doi: 10.1002/ijc.25242. Review
34. Sung J, Wang Y, Chandrasekaran S, Witten DM, Price ND. **Molecular signatures from omics data: from chaos to consensus**. Biotechnol J. 2012 Aug;7(8):946-57.doi: 10.1002/biot.201100305. Epub 2012 Apr 23. Review.
35. Ames B. N., **Dietary carcinogens and anticarcinogens. Science**, (1983), 231, 1256–1264
